STAT 300 Chapter 1

Zepu Zhang
January 20, 2012

1 Some basic concepts

1.1 Basic entities in a statistical study

population: the collection of all the objects of interest.

sample: a subset of the population.

Example Population: all 1st year undergrad students at UAF. Sample: sit at Wood Center bus stop for 1 hour and pick all the 1st year undergrad UAF students you see there.

1.2 Probability vs statistics

"Probability" is firmly a branch of mathematics. "Statistics" is less so in the opinion of some. Many probability books are just like other math books in that you see no actual, observed/measured data or specific numbers in them, it's full of math symbols and formulas; in statistics books you usually see actual data. The distinction between probability and statistics may be understood informally as follows.

What probability does:
reason from the population to the sample—with properties of the population assumed known, study the behavior of a sample from the population.

Example Suppose 10% of the drivers on Fairbanks streets do not wear seatbelt and these drivers are perfectly mixed into all drivers. A police officer stands at a typical spot on the street and checks 3 drivers in a row. What is the probability that all 3 drivers wear seat belts? What is the probability that exactly 2 of the 3 drivers wear seat belts?

What statistics does:
(1) provide guidance to collecting data (sampling tech-

niques, experimental design);
(2) describe and present data (descriptive statistics);
(3) infer about the population (inferential statistics).

Some basic tools of numerical summaries (descriptions) and graphical presentations apply to both a population and a sample. Some others apply to one but not the other.

Inferential statistics: reasons from the sample to the population—with a sample available, draw conclusions about certain characteristics of the population.

Note that 'probability' and 'inferential statistics' work in opposite directions.

Example Assume perfect mixing as above. Suppose an officer checked 5 drivers in a row and found 2 of them not wearing seat belts. What is your guess of the proportion of all Fairbanks drivers who do not wear seat belts?

2 Measures of location

Here "location" means "average" magnitude, or "center".

Sample mean, $\overline{x}$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

n: sample size.

Sample median, $\tilde{x}$

$$\tilde{x} = \begin{cases} \text{the middle value,} & n \text{ odd} \\ \text{average of the two middle values,} & n \text{ even} \end{cases}$$

Example Ex. 1.13, p. 27 (7th ed) or Ex. 1.15, p 30 (8th ed).

> Mean is sensitive to outlying values, whereas median is not—median is a "robust" measure of location. Mathematically, however, mean is much easier to work with than median.

Percentiles, quartiles: describes not necessarily the central value, but the value at a certain relative standing, for example, the 30th percentile is the value such that

30% of the data are below it and 70% are above it. The three quartiles—lower quartile, median, upper quartile—divide the data into 4 equal parts.

Trimmed means: discard some extreme values (e.g. the smallest and the largest) and average the rest.

The preceding discussions are in terms of a "sample" or "data set". To obtain these quantities of a "population", we need knowledge of the entire population or may estimate them based on a sample.

Population mean, μ
population median, $\tilde{\mu}$

3 Measures of variability

Example Fig. 1.17, p. 31 (7th ed) or Fig. 1.19, p. 35 (8th ed).

Range is the simplest measure of variability.

Inter-quartile range (IQR): upper quartile minus lower quartile.

More often and more informatively, we want to characterize the typical deviation from the 'center'.

Mean absolute deviation (from the mean)

median absolute deviation (from the median) (more robust)

Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2$$

$n-1$ is called the "degrees of freedom".

Note Why divide by $n-1$ instead of n? See "Motivation for s^2" in Chap. 1.4. But don't worry about it, at least for now.

Note The summation $\sum_{i=1}^{n}(x_i - \overline{x})^2$ is often denoted by S_{xx}.

Sample standard deviation

$$s = \sqrt{s^2}$$

Example Ex. 1.15, p. 33 (7th ed) or Ex. 1.17, p. 36 (8th ed).

Computing formula

$$\frac{\sum (x_i - \overline{x})^2}{n} = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2$$

In words,

(mean squared deviation) = (mean square)−(squared mean)

Note s^2 is a simple variation to this: it's the left-hand side multiplied by $n/(n-1)$.

Exercise Prove this relation.

The preceding discussions are in terms of a "sample" or "data set". To obtain these quantities of a "population", we need knowledge of the entire population or may estimate them based on a sample.

Population variance, σ^2
population standard deviation, σ

4 Graphics

4.1 Histogram

1. For discrete data: plot the frequency (or relative frequency) of occurrence of each value.

   Horizontal scale is "value" of the variable; vertical scale is "frequency" or "count" or "relative frequency".

2. For continuous data, with equal class width: plot the frequency (or relative frequency) of occurrence in each class. Stick with a boundary rule, e.g. left boundary rule—$[a, b)$ is one class.

   Horizontal scale is "value" of the variable; vertical scale is "frequency" or "relative frequency".

   There should be no gaps between the bars (b/c there is no gap in the value range you are plotting).

Example Ex. 1.9, p. 16 (7th ed) or Ex. 1.10, p. 18 (8th ed).

3. For continuous data, with unequal class width: plot on density scale. The area of each class rectangle is the relative frequency of occurrence of that class, hence the height is the density, that is, $\frac{\text{relative freq}}{\text{class width}}$.

> Horizontal scale is “value” (or the variable); vertical scale is “density”.

$$\text{class height} = \frac{\text{relative freq}}{\text{class width}}$$

$$\text{class area} = \text{class width} \times \text{class height} = \text{relative freq}$$

$$\text{total area of bars} = 1$$

> This type of graphs is good when the value range of the dataset stretches far out, especially to one side. If one uses equal class widths, the class width has to be large (in order to show the large value range), and this will hide details in the high-frequency range.

Example Ex. 1.10, p. 17 (7th ed) or Ex. 1.11, p. 20 (8th ed).

There are no hard-and-fast rules for the choice of the number of classes, the class widths, and the class boundaries. However, these choices may have big influences on the appearance of the graph.

> Some practical advice: use 5–20 classes, or about $\sqrt{\text{number of observations}}$ classes.

Histogram is good for showing the shape of the distribution, highlighting properties such as symmetry or skewness, or modality (i.e. number of peaks).

Example Ex. 1.12, p. 21 (8th ed).

4.2 Boxplots

Concise; very informative. Good for showing symmetry and outliers.

Box plots are also called “box and whisker plots”.

The two edges of the box indicate the lower and upper quartiles. A line inside the box indicates the median. Outside of the box extend two whiskers. There are a number of variations as to how far the whiskers extend to, for example

- The whiskers extend to the minimum and maximum values in the data set.
- The whiskers extend to the smallest value within 1.5 IQR of the lower quartile and the largest value within 1.5 IQR of the upper quartile. Data points outside the range depicted by the whiskers are considered “outliers” or “extremes”, and are plotted individually.

Note There is no universal definition of outliers or extremes. The definition given on page 37 (7th ed) or 40 (8th ed) is just an example.

> Some statistical packages have options for adjusting whether to show outliers, what values are considered extremes, and so on.

five-point summary: min, lower quartile, median, upper quartile, max.

Example Ex. 1.17, p. 36 (7th ed) or Ex. 1.19, p. 40 (8th ed).

Example Ex. 1.18, p. 37 (7th ed) or Ex. 1.20, p. 41 (8th ed).

Side-by-side boxplots are good for comparing several data sets in terms of location, spread, symmetry, outliers, etc.

> Such “comparative boxplots” are used only if the data sets are comparable. (Same nature, same unit, more or less comparable value range.)

Example Ex. 1.19, p. 38 (7th ed) or Ex. 1.21, p. 42 (8th ed).

Recommended reading: the “box plot” entry on Wikipedia.

5 Useful R functions

```
min, max, sort

sqrt, ^

mean, median, quantile, var, sd

plot, hist, boxplot
```